Investigating the Impact of Different Pivot Languages on Translation Quality

Longhui Zou lzou4@kent.edu
Ali Saeedi asaeedi@kent.edu
Michael Carl mcarl6@kent.edu
Modern and Classical Language Studies, Kent State University, Kent, USA

Abstract

Translating via an intermediate pivot language is a common practice, but the impact of the pivot language on the quality of the final translation has not often been investigated. In order to compare the effect of different pivots, we back-translate 41 English source segments via various intermediate channels (Arabic, Chinese and monolingual paraphrasing) into English. We compare the 912 English back-translations of the 41 original English segments using manual evaluation, as well as COMET and various incarnations of BLEU. We compare human from-scratch back-translations with MT back-translations and monolingual paraphrasing. A variation of BLEU (Cum-2) seems to better correlate with our manual evaluation than COMET and the conventional BLEU Cum-4, but a fine-grained qualitative analysis reveals that differences between different pivot languages (Arabic and Chinese) are not captured by the automatized TQA measures.

1 Introduction

Translation via a pivot language has been a common practice for a long time. For instance, the preservation of ancient Greek ideas is a major contribution of Islamic civilization via Arabic as a pivot language. Much of Aristotle's original work in Old Greek is preserved to us through Muslim scholars who translated the ancient-Greek scripts into Arabic which was then later translated into Latin, and from there into various other languages. Still today, pivot translation is an important technique mainly due to a lack of available direct translators. The availability of translators who know two (or more) languages becomes increasingly limited as the number of speakers in those languages decreases. It is, therefore, in particular, translation across smaller languages which requires translation via another, usually a more common language. Thus, from the more than 4000 languages in the world that have developed a writing system [1], translators will be available for only a very tiny fraction of the 16 million or so possible language combinations. However, translations into (or out of) the 'big' languages — such as English, French, Spanish, Russian, or Arabic — might be more easily available. Similarly, there are 552 language pairs for the 24 official European languages but it might not always be possible to find translators for all of these combinations. As a work-around, often English, French, or Spanish are used as an intermediate language in the EU.

While pivot translation is commonly used for written and spoken language (e.g., Interpretation), not much work exists that assesses the impact of the intermediate language on the translation quality. Pieta (2019) indicates that translation studies researchers' interest in pivot

[1] https://www.ethnologue.com/

translation has grown since the mid-2010s. The first research that focuses on pivot translation, however, can be traced back to 1963 which was in regard to literary works (Zaborov, 1963). Zaborov's work reflects the Soviets authorial control over book translation by requiring to translate any foreign book into Russian before it can be translated into other languages (Pieta, 2019). Translation studies' trend of literature-oriented research focusing on pivot translation carried on through the seventies, eighties and nineties of the twentieth century (Radó, 1975; Toury, 1988; aus zweiter Hand, 1984; DURISIN, 1991; Kurtz and Pöhlker, 1999). Starting from 1999 onward, pivot translation research expanded to include two other areas in which translation via an intermediate language is considered a common practice, namely interpreting and audiovisual translation (Gambier, 2003; Zilberdik, 2004; Shlesinger, 2010). More recently, pivot translation is getting more popular in more areas of research. Liu et al. (2018) for instance, review the applicability of pivot MT systems and recommend incorporating "quality estimation and/or automatic/human post-editing to the intermediate translation of the pivot language" (p. 10). Most recently, O'Hagan (2022) investigates the challenges, and implications of the use of English pivot translation in game localization.

The choice of the pivot language is often based on the available human (and/or electronic) resources, but the quality of the final translation depends crucially on the quality of the pivot language. If there is a mistake or ambiguity in the pivot translation, the source meaning might be erroneously or incompletely reproduced in the target. The pivot language might be lacking (linguistic) constructions and possibilities that the source language has and, therefore, be incorrectly recovered from the pivot language. The pivot language might also favor interpretations that lead to incorrect conclusions in the target. As compared to direct translation, pivot translation proceeds in two step (1: source-to-pivot and 2:pivot-to-target), each of which filters or amplifies the linguistic signal in specific ways.

In this study, we use back-translation as a method to assess the impact of different pivot languages in translation. We choose the source and the target to be the same language (English), and we select two quite different pivot languages, Chinese and Arabic. Back-translation into English via two different intermediate languages allows us to clearly assess the impact of the pivot language, since any divergence between the source and the target can be attributed to the intermediate language. We triangulate using monolingual paraphrasing as a tool with which back-translations are compared.

Section 2 provides a detailed description of the different datasets and their collection processes. In Section 3, we explain the different translation quality assessment methods we used. We describe our manual evaluation design and use its result as a reference for the results of the two automatic evaluation metrics we incorporate (BLEU and COMET). Then, we draw quantitative and qualitative comparisons among the three assessment metrics' results. In Section 4, we present a statistically backed discussion of the influence of pivot languages on human translation quality in our datasets. We follow this discussion with an in-depth qualitative observations from our datasets in the light of normalization, priming, and shining-through. Section 5 gives a summary and conclusions and states future endeavors.

2 Experimental Design

This study compares English back-translations via Arabic and Chinese pivot languages and monolingual English paraphrasing on the segment level. We generated data for from-scratch back-translation (HT) via Arabic (AR), Chinese (ZH), via monolingual English paraphrasing (PH), as well as machine back-translation (MT).

A total number of 41 English source segments were first machine translated and post-edited by professional translators into Arabic and Chinese. These Arabic and Chinese translations served as pivot translation. Subsequently, the Arabic pivot translations were then back-translated into English by 8 translators (AR) - with Arabic as their first language (L1) and English as their second language (L2). The Chinese pivot translations were back-translated into English by 4 translators (ZH) - with Chinese as their L1 and English as their L2. For these human from-scratch back-translations, we collected behavioral data, eye-tracking and key logging.

As described in (Saeedi, 2021), we also used two neural machine translation (NMT) systems (i.e., Bing and Google Translate) to generate English NMT back-translations (MT) via the Arabic and Chinese pivot translations. In addition, 8 computer sciences graduate students with English as their L2 produced monolingual paraphrases of the original English segments (EN). A total of 912 translated segments were generated, consisting of: AR 328 segments, ZH 164 segments, MT also 164 segments, from which 82 segments for Arabic (MT-AR) and 82 segments for Chinese (MT-ZH), and EN 256 segments. Figure 1 illustrates this data collection process. It shows how we set up the Translation Quality Assessment (TQA) for three different translation tasks (HT, MT, PH). [2]

The data was processed in the CRITT Translation Process Research Database (Carl et al., 2016, CRITT TPR-DB), which includes manual word-alignment of the 912 segments. The quality of all translated segments were manually evaluated, as well as automatic assessment (i.e., BLEU and COMET). These assessment results were utilized as references for the translation quality.

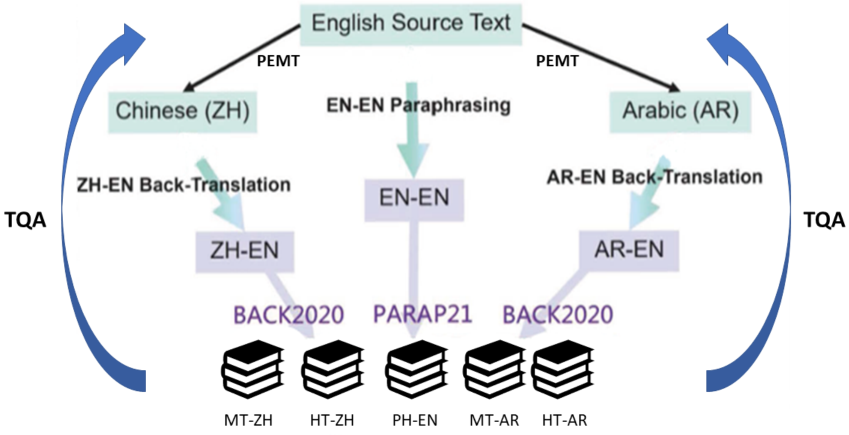

Figure 1: Collection Process of BACK2020 and PARAP21 Datasets

3 Translation Quality Assessment

3.1 Manual evaluation

Manual evaluation is often used as a gold-standard reference to which the performance of automatic metrics are gauged (Papineni et al., 2002; Rei et al., 2020). Several studies proposed criteria for manual evaluation, such as accuracy and fluency (White and O'Connell, 1994; Koehn

[2] MultiLing was used as English source texts (https://sites.google.com/site/centretranslationinnovation/tpr-db/public-studies). Data of paraphrases are gathered in the study PARAP21 while the back-translations are available as BACK2020 in the CRITT TPR-DB.

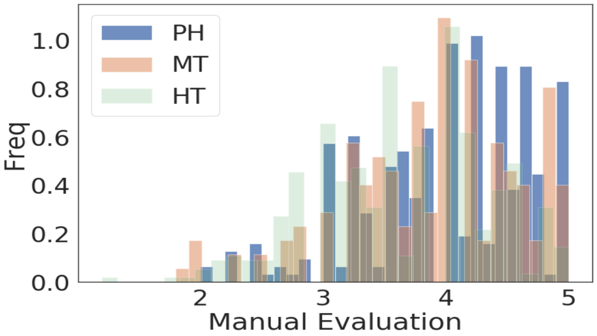

Figure 2: Distribution of Manual Evaluation Scores for Different Translation Tasks: Paraphrasing (PH), Machine Translation (MT), human from-scratch back-translation (HT)

and Monz, 2006; Graham et al., 2015; Barrault et al., 2019; Popović, 2020; Zou et al., 2021). In our study, we select adequacy (i.e., accuracy and fidelity) criteria in view of the fact that accuracy errors are the most severe and often most difficult to detect (White and O'Connell, 1994; Callison-Burch, 2007; Dorr et al., 2010).

Twenty fluent English speakers were recruited as raters, and the evaluation was carried out by rating the (English) back-translations and paraphrasing segments against the (English) source segments according to a likert scale (see Appendix A). Each of the 41 source segments was shown to the raters with 5 candidate translations, and each of the translation segments was rated by 5 raters in different permutations. The inter-rater agreement among all raters was then calculated using the weighted Fleiss's Kappa metric, which showed a good overall agreement of 0.67 (McHugh, 2012).

As a gold standard for segment quality, we used the average manual evaluation score. The distribution of the average manual evaluation per segment for the three translation tasks (i.e., PH, MT, HT) is shown in Figure 2. Evaluators gave overall best scores for PH (μ=3.97, SD=0.68), followed by MT (μ=3.90, SD=0.70), and somewhat less scores to HT (μ=3.64, SD=0.70) in our experiment.

3.2 BLEU

The Bilingual Evaluation Understudy (BLEU) is perhaps the most commonly used automatic metric in TQA research (Doddington, 2002; Dorr et al., 2011; Moorkens et al., 2018). BLEU produces a score between 0 and 1, based on a precision measure that compares n-grams in candidate translations to matching n-grams in reference translations. Specifying the weighting of different n-grams in the calculation of the BLEU score allows for the formation of different types of BLEU scores including individual and cumulative scores. The individual n-gram BLEU scores evaluate the matching grams between the candidate translations and the reference text independently. The cumulative n-gram BLEU scores (referred to as Cum-n) calculate "individual n-gram scores at all orders from 1 to n" and weigh them "by calculating the weighted geometric mean" (Brownlee, 2017). The cumulative 4-gram BLEU score (Cum-4) is the default calculated score for sentence-level or whole-text-level scores (Hailu et al., 2020).

It seems that TQA research seldom delves into different weights of BLEU scores and how they affect the assessment results. We used the `sentence_bleu` function in python to investigate how different configurations of BLEU scores correlate with our manual gold standard evaluation. We calculated the correlation between the BLEU scores and the average manual evaluation for each segment. As we can see from Table 1, 1-gram and Cum-2 scores

	1-gram	2-gram	3-gram	4-gram	Cum-2	Cum-3	Cum-4	COMET	Manual
1-gram	1.0	0.9	0.82	0.73	0.94	0.88	0.81	0.53	0.43
2-gram	0.9	1.0	0.96	0.88	0.98	0.98	0.93	0.46	0.4
3-gram	0.82	0.96	1.0	0.96	0.91	0.97	0.97	0.4	0.37
4-gram	0.73	0.88	0.96	1.0	0.82	0.88	0.95	0.33	0.34
Cum-2	0.94	0.98	0.91	0.82	1.0	0.96	0.89	0.49	0.41
Cum-3	0.88	0.98	0.97	0.88	0.96	1.0	0.94	0.45	0.38
Cum-4	0.81	0.93	0.97	0.95	0.89	0.94	1.0	0.38	0.36
COMET	0.53	0.46	0.4	0.33	0.49	0.45	0.38	1.0	0.37
Manual	0.43	0.4	0.37	0.34	0.41	0.38	0.36	0.37	1.0

All correlations are significant with p <0.01

Table 1: Pearson Correlation Between Different Weights of BLEU Scores, COMET, and Manual Evaluation

correlate best with our manual evaluation results. The correlation coefficients show a moderate relationship between 1-gram (r=0.43), Cum-2 (r=0.41) scores, and manual evaluation (Schober et al., 2018). Given these results, we take it that Cum-2 may be a better assessment method than the commonly used Cum-4. Even though 1-gram provides even better correlation with the human gold-standard assessment, we rule out uni-grams as a viable automatic assessment method as it does not take into consideration any collocational information in the evaluation. Thus we use Cum-2 scores as our selected BLEU weight for segment-level quality assessment. [3]

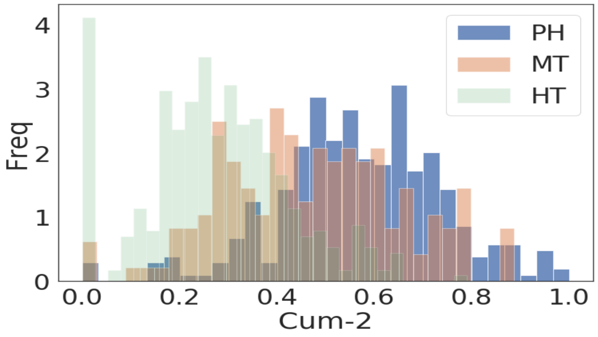

Figure 3: Distribution of BLEU Scores for PH, MT, and HT

We also compare the distributions for Cum-2 scores for the 912 translation segments across the three tasks (HT, MT, PH). As can be gathered from Figure 3, Cum-2 apparently does discriminate between the three translation tasks. Paraphrasing (PH) has overall highest Cum-2 scores (μ=0.57, SD=0.17), followed by MT (μ=0.47, SD=0.19), while human from-scratch back-translation (HT) receives the lowest scores (μ=0.27, SD=0.15). Note that this Cum-2 ranking coincides with the manual evaluation, as in Figure 2, although the discrimination is not as strong in our gold standard.

[3] While larger n-gram may have been useful for earlier MT output to assess fluency issues, shorter n-grams models may better capture translation accuracy. However, with increased quality of recent (N)MT, the main translation problems are due to lack of accuracy.

3.3 COMET

COMET is a neural framework for machine translation evaluation. It can be used to "help evaluate and predict the quality of machine-generated translations for many different languages" (Lavie, 2020). It makes use of word embeddings, which are real-valued vector spaces that encode the meaning (i.e. usage) of the word in context, assuming that words closer in the vector space are expected to be similar in meaning (Teller, 2000). Within COMET, word embeddings "are then passed through a pooling layer to create a sentence embedding for each segment. Finally, the resulting sentence embeddings are combined and concatenated into one single vector that is passed to a feed-forward regressor" (Rei et al., 2020, p. 3). COMET is supposed to better deal with synonymous words, as they are used in similar contexts and thus assigned similar weights. COMET is still a relatively new and understudied automatic assessment metric in TQA research as compared to BLEU.

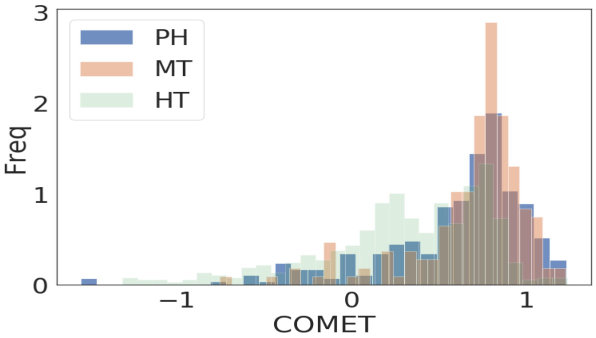

Figure 4: Distribution of COMET Scores for PH, MT, and HT

As illustrated in Table 1, both COMET and manual evaluation correlate best with Cum-2. However, Cum-2 better correlates with our manual evaluation than COMET and the conventional Cum-4. Similar to manual evaluation (section 2) and the BLEU score (section 3), we compare the distribution of COMET scores per segment for the three translation tasks (i.e., PH, MT, HT), as shown in Figure 4. In contrast to the gold standard and Cum-2, COMET gives (on average) highest scores for MT (μ=0.68, SD=0.33) followed by HT (μ=0.59, SD=0.43), then PH (μ=0.30, SD=0.48).

3.4 Comparing Evaluation Metrics

In this section we look at results of the different evaluation methods on a more granular level. Table 2 provides an example that assesses differences between the three evaluation metrics against candidate translations from HT, MT, and PH.

For the HT translation, the Cum-2 score of 0.25 is slightly lower than the Cum-2 average (μ=0.27) for this task, which is likely due to the lack in overlap of uni- and bi-grams between the reference and the candidate translation.

The COMET score (0.68) and the manual evaluation (4.0) for this translation are, in contrast, above their average HT scores of μ=0.30 and μ=3.64 respectively. An explanation for this different assessment may be that COMET and manual evaluation account for semantic similarities rather than the similarity of the words' surface forms. In section 4 we argue that back-translations are less literal than paraphrases or MT (see also Appendix B). Thus, the words *possibly* and *for* in the HT translation can be seen synonymous respectively for *could* and *to* of the reference.

The BLEU Cum-2 scores for other translations (MT and PH) are clearly above the task

Reference	All of them could be considered a burden to hospital staff.			
Task	**Candidate Translation Segment**	**Cum-2**	**COMET**	**Manual**
HT	These victims were possibly considered as the burdens for the hospital staff.	0.25	0.68	4.0
MT	Each of them could be considered a burden on the hospital staff.	0.72	0.74	4.8
PH	He considered all of the a burden to hospital staff.	0.64	0.81	3.8

Table 2: Quality Assessment Scores for Example (1) Among Different Tasks of Translation

average (μ=0.47 and μ=0.57, respectively) which may be due to a larger overlap in word forms. All COMET scores of the translations in Table 2 rank above the task averages. Only the manual evaluation sore (3.8) for the paraphrase is below the task average (μ=3.97). A value of 3.8 falls under the description "some meaning is retained" (see Appendix A). This somewhat lower ranking of the paraphrase can be explained by the typo introduced, *the* instead *them*, and the omission of *could*, both of which does not seem to bother COMET and Cum-2 much.

4 Impact of Different Pivot Languages on Human Translation

In this section we look into differences between paraphrasing (EN) and different human translations via the pivot languages (AR and ZH).

4.1 Distribution of Quality Scores

All three evaluation methods provide relatively higher ratings for paraphrasing (EN) as compared to the Arabic and Chinese back-translations. For manual evaluation, the averages for English paraphrasing (EN, μ=3.97) are significantly higher (p=9.28e-10<.01) than for Chinese (ZH, μ=3.66) and for Arabic (AR, μ=3.63). Moreover, for all three evaluation methods, the distributions of AR and ZH are more similar while EN is set apart. This similarity of distributions can be observed in Figure 5 for all three evaluation methods, manual evaluation, BLEU (Cum-2) and COMET.

Higher scores for paraphrasing may be attributed to the priming effect of the English source language. Stronger priming effects can be expected if the prime is more similar to the target. Carl and Schaeffer (2017) discussed priming effects in post-editing (PEMT) and from-scratch translation. They found that "PEMT produces more literal translations than from-scratch translation" (p. 53), due to the fact that MT output is in almost every aspect closer to the final translation than the ST. A similar effect may be expected for monolingual paraphrasing which resembles PEMT, in some sense, as the prime and the target are in same language in both cases. Monolingual paraphrasing might thus render the target segments more literal as compared to the back-translations. The higher degree of literality — in turn — may explain the higher quality ratings since there may be less variation in surface forms which are closer to the source.

4.2 Translation Variation

In this section we look into the translation variation produced in the different (EN, AR, ZH) channels. As Table 3 shows, back-translations seem to introduce more variation, while English paraphrasing yields more literal renditions. Table 3 plots an English reference sentence and different ways in which the ST word *nomadic* was rendered via paraphrasing and the back-translations. While there is much variation in the back-translations, all 8 monolingual paraphrases make use of different derivations of the same lexeme *nomad*. The literal rendition,

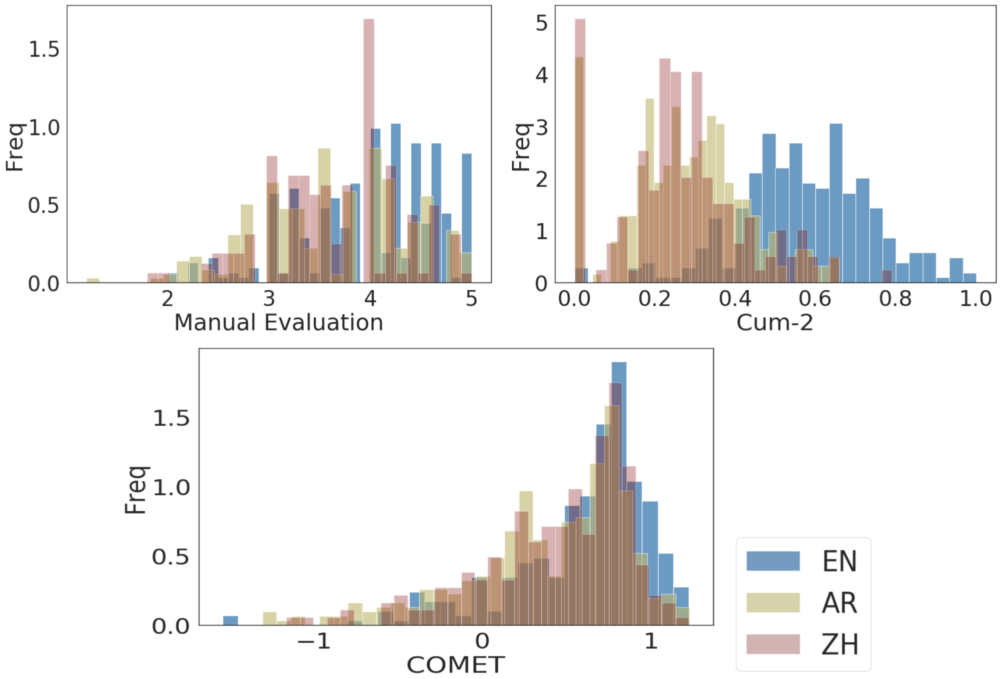

Figure 5: Distribution of manual evaluation scores, BLEU (Cum-2) and COMET for Paraphrasing (EN), back-translation via Arabic (AR) and Chinese (ZH). EN has highest scores and is more clearly separated from the back-translations for all metrics.

nomadic, occurs in 5 out of 8 instances (62.5%). The other 3 instances only change the part of speech, *nomads*, or grammatical number, *nomadics*, albeit incorrectly. This observation corroborates our priming assumption, which suggests that stronger priming effects in monolingual paraphrasing results in more literal translations (see also appendix B).

Carl and Schaeffer (2017) found that there is “more lexical variation in from-scratch translations than in post-editing” (p. 55). In addition, back-translations draw from one extra step of forward translation, which has the potential to introduce more synonymous in the translation. The word *bedouin* is used in 5 out of 8 (62.5%) different back-translations from Arabic with different spellings. This is most likely due to the Arabic pivot translation *min al-mujtama’at al-badawiia al-raHala*, which translates into “of the Bedouin nomadic societies”. The word *badawī* in Arabic refers to the nomadic Arab of the desert.

Amponsah-Kaakyire et al. (2021) state that “information about native language and qualifications of the translator is [...] relevant” when analyzing multilingual corpora to study translationese including “language independent characteristics like simplification, normalization, explicitation and avoiding repetitions ... [and] language-pair specific features” like “shining-through of source language patterns in target text” (p. 1). Taking into account that the Arabic pivot translation was produced by a professional translator into his L1, we assume that the word choice “based on [their] subjectivity, is a part of normalization process” (Imjidee and Kwee, 2020, p. 1). The 5 instances of occurrences of *bedouin* in the back-translations suggest thus a shining-through of Arabic pivot translation.

Reference	The majority of hunter-gatherer societies are *nomadic*.		
Pivot	**TT Token(s)**	**Frequency**	**Percentage**
AR	nomadic	3	37.5%
	bedouin nomads	1	12.5%
	bedouin, nomad communities	1	12.5%
	once of the Beduin traveller comunities	1	12.5%
	types of the Beduin traveller communities	1	12.5%
	beduins	1	12.5%
ZH	nomadic	2	50.0%
	moving	1	25.0%
	drift from place to place	1	25.0%
EN	nomadic	5	62.5%
	nomadics	2	25.0%
	nomads	1	12.5%

Table 3: Human Translations for “nomadic” from Pivot Languages and Monolingual Paraphrasing

4.3 Shining through

Teich (2003) stipulates that “what makes translation different from original texts in the same language as the target language is that the source language shines through in translations” (p. 219). Lapshinova-Koltunski (2015) hypothesize that the languages with a higher status tend to ‘shine through’ more often assuming that in “translations from English, we would probably observe more “shining through [...] as English has the highest world language status (p. 97). From the quantitative analysis in Figure 5, we see that the influence of our pivot languages (AR and ZH) does not seem to have a measurable effect on translation quality, despite that fact that there are qualitatively very different translation variations. We assume this is due to the cultural differences in these two languages.

Reference	[...] Norris *disliked* working with old people.		
Pivot	**TT Token(s)**	**Frequency**	**Percentage**
AR	hated	2	25.0%
	hated to	1	12.5%
	hatred of	1	12.5%
	hates	1	12.5%
	had got to hate	1	12.5%
	did not like	1	12.5%
	disliking	1	12.5%
ZH	doesn’t like	2	50.0%
	do not like	1	25.0%
	did not like to	1	25.0%

Table 4: Back-translations for “disliked” via Arabic and Chinese Pivot Languages

In view of this, we further zoom in to the different AR and ZH translations. Table 4 shows (a part of) a reference sentence with 12 AR and ZH back-translations and the different ways in which the token *disliked*, was reproduced. The majority of the Arabic participants — 6 out of 8 (75%) — translated into “hate”, only 2 of them (25%) translated into the equivalent meaning *did not like*, or *disliked*. On the contrary, all the Chinese participants translated into some versions of *not like*. We see this as another example of shining through. From the last two examples, we see that different source languages shine through the target text differently since shining-through is a language-pair specific feature.

5 Conclusion

This study investigates the quality of translations via different pivot languages. In order to allow for seamless comparison of the source and the target, it compares human back-translation, neural machine back-translation, and monolingual paraphrasing from English via Arabic and Chinese back into English. Six short English texts (together 41 sentences) were translated into Arabic and Chinese. The two sets of Arabic and Chinese texts were then back-translated into English by 8 Arabic and 4 Chinese translation students respectively. In addition, we also produced back-translations with two NMT systems, Bing and Google Translate, and the English original texts were also paraphrased by 8 computer sciences students. This amounts to 25 English versions: 1 English original, 8 Arabic and 4 Chinese human back-translations, 2 NMT back-translations from Arabic and 2 from Chinese, and 8 (monolingual) paraphraes. However, as some segments were not translated, paraphrased or lost due to software errors, we were left with 912 translated segments (from potentially 24*41 = 984).

We assessed the quality of the 24 reproduced versions (912 translated segments) on a segment level by comparing each of them with the English original, using two automatic measures (BLEU and COMET) as well as manual evaluation. The manual evaluation was based on a Likert scale (see Appendix A) in which 20 fluent English speakers assessed the similarity between the original and the reproduced versions. Each segment was independently rated by at least five evaluators with good agreement (weighted Fleiss's Kappa of 0.67). We took the average score of the manual evaluations as a gold standard. We also experimented with different weights for various BLEU configurations. Our findings indicate that:

- Monolingual paraphrasing has the best scores across our three evaluation methods.
- NMT back-translations achieved similar quality ratings compared to the human back-translations.
- The BLEU Cum-2 measure correlates better with our (averaged) manual gold evaluations than conventional BLEU Cum-4 and COMET.
- Despite qualitative differences, our automatic metrics cannot separate between different pivot languages (AR and ZH).

We explain the higher scores of monolingual paraphrasing (as compared to bilingual paraphrasing) through stronger priming effect, which results in more literal renderings. In contrast, normalization processes effects could be observed in the forward translation when generating the pivot translation, which we explained as cultural differences in the pivot language, while shining-through effects could be observed in the back-translation. Both phenomena may be factors which results in back-translations to show more lexical variation than paraphrasing.

Further studies can be conducted using other automatic metrics and similar methods applied in this study. Further research can also include more pivot languages and NMT systems. Furthermore, the collected behavioral data of eye tracking and key logging of the human from-scratch back-translations can be utilized for purposes of triangulation and also in future related research.

References

Amponsah-Kaakyire, K., Pylypenko, D., España-Bonet, C., and van Genabith, J. (2021). Do not rely on relay translations: Multilingual parallel direct europarl. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 1–7.

aus zweiter Hand, Ü. (1984). Rezeptionsvorgänge in der europäischen literatur vom 14. bis zum 18. jahrhundert.

Barrault, L., Bojar, O., Costa-Jussa, M. R., Federmann, C., Fishel, M., and Graham, Y. (2019). Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of ACL*. Association for Computational Linguistics (ACL).

Brownlee, J. (2017). A gentle introduction to calculating the bleu score for text in python. *Section on Deep Language Processing. Accessed August*, 20:2019.

Callison-Burch, C. (2007). *Paraphrasing and translation*. PhD thesis, University of Edinburgh Edinburgh.

Carl, M., Bangalore, S., and Schaeffer, M. (2016). New directions in empirical translation process research. *Heidelberg: Springer International Publishing Switzerland. doi*, 10:978–3.

Carl, M. and Schaeffer, M. J. (2017). Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *HERMES-Journal of Language and Communication in Business (56)*, pages 43–57.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Dorr, B., Olive, J., McCary, J., and Christianson, C. (2011). Machine translation evaluation and optimization. In *Handbook of natural language processing and machine translation*, pages 745–843. Springer.

Dorr, B. J., Passonneau, R. J., Farwell, D., Green, R., Habash, N., Helmreich, S., Hovy, E., Levin, L., Miller, K. J., Mitamura, T., et al. (2010). Interlingual annotation of parallel text corpora: a new framework for annotation and evaluation. *Natural Language Engineering*, 16(3):197–243.

DURISIN, D. (1991). Artistic translation in the interliterary process. *TTR Studies in the Text and its Transformations*, 4(1):115–127.

Gambier, Y. (2003). Working with relay: An old story and a new challenge. *Speaking in tongues: Language across contexts and users*, 47:66.

Graham, Y., Baldwin, T., and Mathur, N. (2015). Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191.

Hailu, T. T., Yu, J., and Fantaye, T. G. (2020). A framework for word embedding based automatic text summarization and evaluation. *Information*, 11(2):78.

Imjidee, N. and Kwee, S. B. (2020). Normalization techniques for translating cultural-specific expressions. *LSP International Journal*, 7(2):1–18.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

Kurtz, J. and Pöhlker, K. (1999). *De L'Un Au Multiple. Traduction Du Chinois Vers Les Langues Européennes/Translation from Chinese Into European Languages.* Les Editions de la MSH.

Lapshinova-Koltunski, E. (2015). Variation in translation: Evidence from corpora. *New directions in corpus-based translation studies*, 1:93.

Lavie, A. (2020). Why we built comet, a new framework and metric for automated machine translation evaluation.

Liu, C.-H., Silva, C. C., Wang, L., and Way, A. (2018). Pivot machine translation using chinese as pivot language. In *China Workshop on Machine Translation*, pages 74–85. Springer.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (2018). Translation quality assessment. *Machine translation: Technologies and applications ser. Cham: Springer International Publishing*, 1:299.

O'Hagan, M. (2022). Indirect translation in game localization as a method of global circulation of digital artefacts: A socio-economic perspective. *Target*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pieta, H. (2019). Indirect translation: Main trends in practice and research. *Slovo.ru: Baltic accent*, 10:21–36.

Popović, M. (2020). Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264.

Radó, G. (1975). Indirect translation. *Babel*, 21(2):51–59.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Saeedi, A. (2021). Comparing backtranslations across different pivot languages and translation modes. In *Proceedings of The international and interdisciplinary conference on Applied Linguistics and Professional Practice (ALAPP)*, page 18.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Shlesinger, M. (2010). Relay interpreting. *Handbook of translation studies*, 1:276–278.

Teich, E. (2003). *Cross-linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*, volume 5. Walter de Gruyter.

Teller, V. (2000). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.

Toury, G. (1988). Translating english literature via german and vice versa: A symptomatic reversal in the history of modern hebrew literature. *Die literarische Übersetzung: Stand und Perspektiven ihrer Erforschung*, pages 139–157.

White, J. S. and O'Connell, T. A. (1994). Evaluation in the arpa machine translation program: 1993 methodology. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Zaborov, P. (1963). 'literatura-posrednik'v istorii russko-zapadnych literaturnych sviazej xviii-xix vv. *InMezhdunarodnye sviazi russkoj literatury. Sbornik statej, edited byM. Alekseev*, pages 64–85.

Zilberdik, N. J. (2004). Relay translation in subtitling. *Perspectives: Studies in translatology*, 12(1):31–55.

Zou, L., Carl, M., Mirzapour, M., Jacquenet, H., and Vieira, L. N. (2021). Ai-based syntactic complexity metrics and sight interpreting performance. In *International Conference on Intelligent Human Computer Interaction*, pages 534–547. Springer.

Appendices

A Description of manual evaluation

For manual evaluation guidelines, we use a Likert scale with the following values:

5	All meaning is retained
4	Most meaning is retained
3	Some meaning is retained
2	Little meaning is retained
1	No meaning is retained

Table 5: Description of Manual Evaluation Metrics

B Segment-wise Target Source Token Ratio

In order to assess a level of literal (i.e., word-for-word translation) vs. free translation for each of the three tasks, we calculate Target/Source Token Ratio (TSR) for each segment, by dividing the number of tokens in the target segment (TokT) by the number of source tokens (TokS). We can see from Figure 6 that HT has the most TSR variation followed by MT and PH. We take higher TSR variation as an indicator for free translation, and low(er) TSR as indicator for more literal translation. According to the TSR measure more literal translations were produced in PH, followed MT, while HT is least literal.

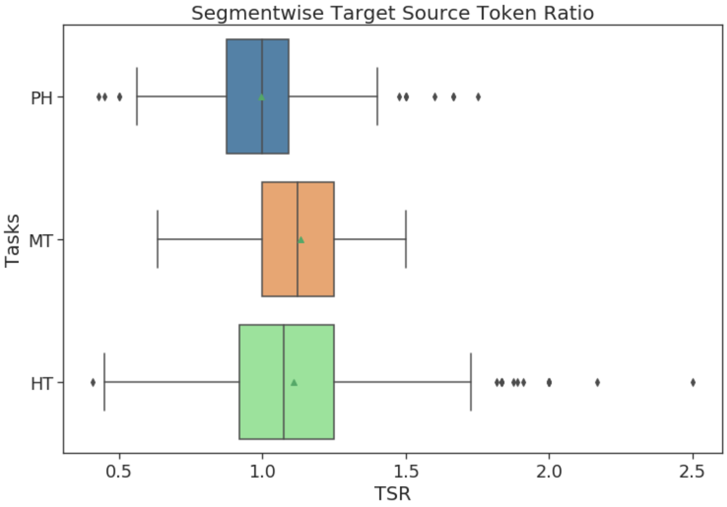

Figure 6: Segment-wise Target Source Token Ratio